Advances in Data Analysis and Classification (2020) 14:513–515
https://doi.org/10.1007/s11634-020-00421-8

EDITORIAL

Editorial for ADAC issue 3 of volume 14 (2020)

Maurizio Vichi[1] · Andrea Cerioli[2] · Hans Kestler[3] · Akinori Okada[4] · Claus Weihs[5]

This issue 3 of volume 14 (2020) of the journal *Advances in Data Analysis and Classification (ADAC)* contains eight articles that deal with decision trees, clustering via point processes, classification with paired covariates, a variant of canonical analysis, topological classification, Chi square decomposition, rank tests for functional data, and group comparisons with the heterogeneous choice model.

Gerhard Tutz contributes the first paper of this ADAC issue with the title "Modelling heterogeneity: On the problem of group comparisons with logistic regression and the potential of the heterogeneous choice model". He investigates the potential of the heterogeneous choice model as a useful tool to analyze the possibly complex effects of explanatory variables and account for interactions in a specific sparse way. Specific advantages of the model are: it can account for effect modifiers, which might represent heterogeneity or uncertainty; several variables can be included in the effect modifying term: it allows for sparse models, most often one obtains a main effects model in the linear predictor. Moreover, a model selection strategy is proposed that can distinguish between effects that are due to heterogeneity and substantial interaction effects. In contrast to the common understanding, the heterogeneous logit model is considered as a model that contains effect modifying terms, which are not necessarily linked to variances, but can also represent other types of heterogeneity in the population.

In the second paper "Is-ClusterMPP: clustering algorithm through point processes and influence space towards high-dimensional data" written by *Khadidja Henni, Pierre-Yves Louis, Brigitte Vannier* and *Ahmed Moussa* a new version of the ClusterMPP, a density-based clustering algorithm using marked point processes and an influence space, is proposed. The new algorithm is an efficient and enhanced version that speeds up the clustering process and is able to boost the detection of adjacent

✉ Maurizio Vichi
maurizio.vichi@uniroma1.it

1 Rome, Italy

2 Parma, Italy

3 Ulm, Germany

4 Tokyo, Japan

5 Dortmund, Germany

clusters with varying densities. It needs less input parameters and uses a parametric statistical model with less parameters that, according to the authors, limits the risk of over-fitting. This improved version of ClusterMPP uses the cardinality of the influence space to reduce the number of parameters, and avoids the need to specify the two parameters Eps and MinPts like in the DBSCAN case. The new approach improves also the detection of clusters with varying densities, because it is highly sensitive to local density changes.

The third article entitled "Sparse classification with paired covariates" is written by *Armin Rauschenberger, Iuliana Ciocănea-Teodorescu, Marianne A. Jonker, Renée X. Menezes* and *Mark A. van de Wiel*, introduces a generalisation of the lasso method for paired covariate settings, named paired lasso. Single response from two high-dimensional covariate sets are proposed. It is often unknown which of the two covariate sets leads to a better prediction, or whether the two covariate sets complement each other. The paired lasso addresses this problem by weighting the covariates to improve the selection from the covariate sets and the covariate pairs. It thereby combines information from both covariate sets and accounts for the paired structure. The paired lasso was tested on more than 2000 classification problems with experimental genomics data, it appeared that for estimating sparse but predictive models, the paired lasso outperforms the standard and the adaptive lasso. The R package *palasso* is available from cran.

The next article on "Connecting the multivariate partial least squares with canonical analysis: A path-following approach" is written by *Lukáš Malec* and *Vladimír Janovský*. They introduce a link between one variant of partial least squares (PLS) and canonical correlation analysis (CCA) for multiple groups, as well as for two groups covered as a special case. A continuation algorithm based on the implicit function theorem is selected, with particular attention paid to potential non-generic points based on real economic data inputs. Both degenerated crossings and multiple eigenvalues are identified on the paths. The theory of Chebyshev polynomials is applied in order to generate novel insights into the phenomenon that is simply generalisable to a variety of other techniques.

The fifth paper is written by *Vasileios Maroulas, Cassie Putman Micucci* and *Adam Spannaus* on "A stable cardinality distance for topological classification". Authors incorporate topological features via persistence diagrams to classify point cloud data arising from materials science. Persistence diagrams are multisets summarizing the connectedness and holes of given data set. A new distance measuring the similarity of persistence diagrams using the cost of matching points and a regularization term corresponding to cardinality differences between diagrams is proposed.

Stability properties of this distance provides theoretical justification for its use in the comparisons of such diagrams.

In the sixth article, the authors *Rosaria Lombardo, Yoshio Takane* and *Eric J. Beh* propose "Familywise decompositions of Pearson's Chi square statistic in the analysis of contingency tables". The authors present a general framework for partitioning Pearson's Chi square statistic under the assumption of complete independence among the variables. This framework is "general", as it applies to contingency tables of any order. The resulting partitions are unique, exact, and can be calculated in closed form, unlike some of their competitors (e.g., the log-likelihood ratio Chi square). With these

partitions, simultaneous tests of marginal and joint probabilities in contingency tables become feasible. This framework accommodates the specification of theoretically driven probabilities as well as the well-known cases in which the marginal probabilities are fixed or estimated from the data. The former allows tests of prescribed marginal probabilities, while the latter allows tests of the associations among variables after eliminating the marginal effects. Mixtures of these two cases are also permitted. Examples are given to illustrate the tests.

In the next paper entitled "Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations", *Alba M. Franco Pereira* and *Rosa E. Lillo* propose new strategies to analyze graphically functional data in the case when there is no total order between the data of the sample, taking into account the information provided by the down-upward partial orderings based on the hypograph and the epigragh indexes. An alternative way to measure centrality in a bunch of curves is obtained by combining the two indexes, thereby an alternative measure of the statistical depth is given. The authors, motivated by the visualization in the plane of the two measures for two functional data samples, propose new methods for testing homogeneity between two groups of functions. The performance of the tests is evaluated through a simulation study and the method is applied to several real data sets.

Finally, in the eighth paper, with the title "Enhancing techniques for learning decision trees from imbalanced data" the authors *Ikram Chaabane, Radhouane Guermazi and Mohamed Hammami* propose a novel adaptation of the decision tree algorithm to imbalanced data situations. Authors provide a new asymmetric entropy measure that best performs essentially in terms of the minority class prediction without sacrificing the majority class prediction. It adjusts the most uncertain class distribution to the a priori class distribution and uses the adjusted distribution in the node splitting-process. Unlike most competitive split criteria, which include only the maximum uncertainty vector in their formula, the proposed entropy is customizable with an adjustable concavity to better comply with the system expectations. The experimental results show significant improvements over various split criteria adapted for imbalanced situations. Furthermore, being combined with sampling strategies and based-ensemble methods, the proposed entropy attains significant enhancements on the minority class prediction, along with a good handling of the data difficulties related to the class imbalance problem.